Multiple phonetically trained-listener comparisons of speech before and after articulatory intervention in two children with repaired submucous cleft palate

Zoe Roxburgh [1], Joanne Cleland[1&2], and James M. Scobbie[1]

[1] Queen Margaret University, [2] University of Strathclyde

zroxburgh@qmu.ac.uk

0131 474 0000

Date of Acceptance: 18th December 2015

Multiple phonetically trained-listener comparisons of speech before and after articulatory intervention in two children with repaired submucous cleft palate

Abstract

In Cleft Palate (CP) assessments based on phonetic transcription are the "gold standard" therapy outcome measure, despite reliability difficulties. Here we propose a novel perceptual evaluation, applied to ultrasound-visual biofeedback (U-VBF) therapy and therapy using visual articulatory models (VAMs) for two children with repaired submucous CP.

Three comparisons were made: post VAM, post U-VBF and overall pre- and post-therapy. Twenty-two phonetically-trained listeners were asked to determine whether pre- or post-therapy recordings sounded "closer to the English target", using their own implicit stored knowledge (prompted via orthographic representation) as a comparison. Results are compared with segment-oriented percent target consonant correct (PTCC) derived from phonetic transcriptions by the authors.

Listener judgements and PTCC suggest that both children made improvements using both VAM and U-VBF. Statistical analysis showed listener agreement across all three comparisons, despite agreement being poor. This perceptual evaluation offers a straightforward method of evaluating the effectiveness of interventions and can be used by phonetically trained or lay listeners.

Keywords: Ultrasound, Visual Articulatory Models, Perceptual Speech Evaluation, Cleft Palate.

Introduction

Cleft lip and palate (CLP) is one of the most common congenital malformations, with a worldwide incidence of 1.2/1000 (Rahimov, Jugessur and Murray, 2012). Due to the resultant structural abnormalities, children are at a high risk of developing speech difficulties (Vallino-Napoli, 2011). Whilst this is routinely managed surgically, Hardin-Jones and Jones (2005) report that the majority of pre-schoolers with palatal repairs (68% of 212 preschool-aged children) still require therapy focused on improving their speech. Consensus on the best types of intervention for treating these articulatory errors is lacking, with a recent systematic review finding little evidence to support any particular technique (Bessell, Sell, Whiting, Roulstone, Albery, Persson, and Ness, 2013). However, of the 17 studies that did meet inclusion criteria, 10 report the results of motor based approaches, suggesting that the professional opinion is that interventions which capitalise on the principles of motor-learning may be appropriate for this client group (Ruscello and Vallino, 2014). Although interventions may employ instrumental techniques (see below), the primary aim of therapy is for clients to develop speech perceptually similar (if not indistinguishable) to their peers (Britton, Albery, Bowden, Harding-Bell, Phippen, and Sell, 2014), and hence perceptual outcomes and perceptual evaluation are key. Kuehn and Moller (2000) suggest that it is this method that has the greatest face validity, with perceptual speech assessment considered a key outcome measure in CLP management (Lohmander and Olsson, 2004; Sell, 2005). In practice, however, phonetic transcription is subjective, particularly in complex speech sound disorders (SSDs) such as those found in CP, which is often associated with low transcriber agreement (Shriberg and Lof 1991). Reliability of narrow transcription can be improved through repeated analysis of recordings augmented with instrumental analysis, plus interactive discussion between transcribers, together giving more consistent (and perhaps more accurate) results than independent live transcription (Amorosa, von Benda, Wagner, and Keck, 1985), but this is expensive and time-consuming.

Perceptual Evaluation of Cleft Palate Speech

To circumvent some of the problems with phonetic transcription, simpler and more generic perceptual experiments have been used to evaluate post-therapy intelligibility/acceptability. Britton et al. (2014) note that perceptual assessment of CP speech should be based on robust listening procedures, that multiple phonetically trained listeners should be used, and that inter- and intra-rater reliability should be included within this robust procedure. Lohmander and Olsson (2004)'s earlier review of perceptual assessments of CP speech found that many of the studies (28 of 88) used only one listener and only eight studies used more than 10 listeners. An interval scale was the most common method of judgement, with phonetic transcriptions only being used in eight studies. It was concluded that many of the studies did not use or report reliability measures. In our view, multiple listener perceptual evaluations of therapy outcomes are more likely to be adopted if they avoid any need for high levels of phonetic training, narrow transcription, and cross-transcriber discussion. An alternative proposal is a holistic comparative judgement between two tokens of recorded speech from different stages in treatment, requiring no skills beyond an ability to make a mutual comparison of words, and no specialised knowledge beyond an intuitive grasp of generally agreed norms and the range of variants that are acceptable in the target language. Preferably such a mutual comparison of spoken forms with each other, in reference to the target, should be undertaken independently by multiple listeners (for logistical simplicity) who can focus in detail on a single speaker in order that the listener can become attuned. Finally, if this approach were to be applied using a bank of regular volunteers, such as the approach used in Ziegler & Zierdt (2008), then the listeners would be already familiar with the general nature of the task.

In this study we present a novel perceptual evaluation which is intended to be easy to use for both clinical researchers and the listeners. We evaluate pre- versus post-therapy versions of a large set of whole words from two single speakers, with phonetically trained listeners primed to focus on the speech sound targeted in therapy. Listeners compare two audio versions of the same word and choose one as being "closer to the English target" on whatever intuitive level they like, with no

specific instructions about the relative importance of phonological, phonetic or prosodic differences between the tokens other than the knowledge of the therapeutic target. A listener's set of judgments might show no significant difference overall between the pre/post therapy sessions, or might show a clear a preference for one over the other. A set of listeners might agree with each other, or come to no shared conclusion.

Finally, by being internally and externally relativistic (combining comparisons between two acoustic tokens and also between both tokens and an implicit set of typical acceptable variants of the target), the procedure is intended to be able to discriminate fine-grained improvements in speech that is both near-target and hence likely to be "correct" in a PTCC transcription, and in speech that is severely disordered and hence likely to be "incorrect", avoiding some effects of phonemic false evaluation (Buckingham and Yule 1987).

Instrumental Techniques

Although we propose that this perceptual evaluation method could be used for any therapy designed to improve segmental accuracy of people with SSDs, we tested the approach using data from a wider study designed to compare two different types of therapies in the cleft palate (CP) population: Ultrasound Visual biofeedback (U-VBF) and Visual Articulatory Model therapy (VAM) (see below). Both are motor-based therapies. U-VBF follows on from the tradition of using electropalatography (EPG) to treat persistent articulatory errors associated with CP, with EPG recommended as an intervention technique by the UK's professional body for Speech and Language Therapists (SLTs) for the last 10 years (RCSLT, 2005). However EPG has several disadvantages for the CP population. Firstly, in CP speech, active compensatory articulations occur due to velopharyngeal insufficiency causing difficulty producing high pressure consonants (Harding and Grunwell, 1998). These compensatory articulations are often characterised by posterior placements not normally found in English, for example pharyngeal or glottal stops (Trost, 1981; Harding and Grunwell, 1998). These types of errors are not imageable with EPG (since it samples only as far back as the juncture of

the hard and soft palate) and are displayed only as an “open pattern”, however the target consonant (at least in English) is imageable, making it possible to use EPG for biofeedback. Secondly, EPG is not suitable for all clients with CP due to requirements for secondary surgery or on-going dental, orthodontic or maxillary input. In contrast, ultrasound tongue imaging (UTI) images from near the tongue tip to virtually the root, with pharyngeal articulations clearly visible. UTI can also be used as a visual biofeedback technique (U-VBF), and a small but growing body of research suggests it is effective in various SSDs (see for example, Preston, Brick, and Landi, 2013; McAllister Byun, Hitchcock, and Swartz, 2014; Cleland, Scobbie and Wrench, 2015). However, to our knowledge, no studies have been undertaken with the CP population. Some preliminary diagnostic work using UTI in CP has been undertaken, with two studies exploring compensatory articulations, for example retraction to posterior placement (Gibbon and Wolters, 2005; Bressmann, Radovanovic, Kulkarni, Klaiman, and Fisher, 2011). Zharkova (2013) also proposes ultrasound-based measurements to analyse the articulation of clients with CP. Bressmann et al. (2011) showed promise for the use of ultrasound in investigating the retracted placement of velar targets and provided additional information on double articulations, common in speakers with CP. Together these studies point clearly in the direction of trialling ultrasound as an intervention technique for this population.

In this study we contrast U-VBF with a motor-based articulatory therapy (a Visual Articulatory Model, VAM) which allows clients to view idealised animations of articulations, but without biofeedback of their own articulations. This allowed us to test our perceptual evaluation on data from multiple time points in the therapeutic process, as is necessary when contrasting two therapies in a single case study design. Whilst there is new evidence of the effectiveness of U-VBF, VAMS are relatively untested. Despite this, they are becoming increasingly popular as a cheap and readily accessible tool for the speech therapy clinic. The current study therefore employs a perceptual evaluation of a pilot comparison of an app, Speech Trainer 3D (Smarty Ears, 2011), with U-VBF in two case studies. Unlike any demonstrations a clinician may give with live ultrasound of a client, animations from Speech Trainer 3D are based on estimations – in fact a stylised animation. However,

in this pilot study we chose to use a commercially available app rather than something more anatomically correct. This gives the study ecological validity by selecting software that is easily available and attractive to clinicians.

Aims

This study proposes a perceptual evaluation of listener judgements comparing pairs of sessions during two blocks of therapeutic intervention. The perceptual evaluation was used to evaluate both U-VBF therapy and motor based therapy using VAM independently of each other, and to evaluate overall improvement from baseline to maintenance. Two children with repaired submucous CP first received a block of VAM therapy using Speech Trainer 3D followed by a block of U-VBF therapy (see below). Analysis of the actual ultrasound data collected during both interventions will be presented in future work. The perceptual evaluation aimed to determine whether there was an improvement in a probe, a set of words not used in treatment. Our hypothesis was that a token of a word recorded later in the therapy timeline would be "closer to the English target" than one recorded earlier. This study also aimed to determine whether the perceptual evaluation reported is a valid tool for evaluating therapy outcomes generally. Our research questions were:

1. Do listeners select time point B (chronologically later in the therapy period) as "closer to the English target" more often than time point A (chronologically earlier in the therapy period)? Three specific comparisons are considered:

    a. VAM (pre/post) Comparison: immediately before and after therapy block one, using VAM
    b. U-VBF (pre/post) Comparison: immediately before and after therapy block two, using U-VBF
    c. BL-M Comparison: baseline (Assessment session 1 prior to any therapy) to maintenance (Assessment session 6: 3 months after both blocks of therapy)

Hypothesis: Listeners will select chronologically later time-points as being "closer to the adult target" if therapy is successful.

2. Are the results from the perceptual evaluation in line with the Percent Target Consonant Correct (PTCC) scores derived from phonetic transcriptions carried out at each assessment?

Hypothesis: Successful therapy will also be shown by rising PTCC scores.

Method

Participants

Speakers

Speakers (participants receiving intervention) were two Scottish males with repaired submucous CP, Andrew and Craig (pseudonyms). Andrew was 9;2 years old and was backing /n/ to palatal or velar placement with suspected double articulations. He had symptoms of velopharyngeal dysfunction (VPD), presenting with audible nasal emission and velopharyngeal friction. He also presented with /s/ distortions. He had previously received extensive therapy to target his production of /n/, with no success. Craig was 6;2 years old and had few high pressure consonants. He was backing /k/ to glottal placement and fronting /g/ to [d] or [n], with suspected double articulations. He also presented with symptoms of VPD, with hypernasal resonance and inconsistent velopharyngeal friction on high pressure consonants. At the time of referral, Craig had not received any input by his SLT to target his production of velars. The focus of previous input had been bilabial consonants and alveolar fricatives Further details on each speaker's error patterns can be found in Roxburgh, Scobbie & Cleland (2015).

Listeners

Twenty-four phonetically trained listeners, three male, 21 female, were recruited from a university in Central Scotland. All listeners had English as a first language, with mixed Scottish, Irish and English accents. Listeners with known speech, language or hearing impairments were excluded from the study. Five listeners were qualified SLTs working at the university, with the remaining 19 being SLT students who had completed phonetics training as part of the undergraduate and postgraduate programmes. Six listeners had previous experience in working with children with CLP. Listeners' level of experience was not equally spread across both speakers; rather listeners were randomly allocated to evaluate a particular speaker. Twelve listeners evaluated Andrew and 12 listeners evaluated Craig, with one listener withdrawing from the study and one listener's data being lost, leaving 10 listeners who evaluated Andrew.

Therapeutic Design

Each child received six assessment/recording sessions and two blocks of therapy, each with eight one-hour therapy sessions. Table 1 provides a schedule for assessment and therapy sessions and outlines the assessments administered within each of the assessment sessions. More detail on the intervention is available in Roxburgh, Scobbie & Cleland (2015).

Insert table 1 about here

Recording Set-up

All assessment sessions were recorded with simultaneous ultrasound, audio and lip-camera. The Research SLT (first author) was blinded to the ultrasound data until assessment 3 so this would not influence the treatment in the VAM condition. Ultrasound was acquired using an Ultrasonix SonixRP machine remotely controlled via Ethernet from a PC running Articulate Assistant Advanced software™ (Articulate Instruments Ltd, 2012) version 2.14 which internally synchronised the ultrasound and audio data. The echo return data was recorded at ~121 frames per second (fps), i.e. ~8ms per frame, with a 135 degree field of view in the mid-sagittal plane.

Baseline Measures

Language measures were taken at baseline. The Core Language Score (CLS) of the Clinical Evaluation of Language Fundamentals 4th Edition (*CELF4*, Semel, Wiig and Secord, 2006) showed that both children fell within the normal range (Andrew CLS 99, Craig CLS 93). The British Picture Vocabulary Scale 3rd Edition (*BPVSIII*, Dunn, Dunn and National Foundation for Educational Research, 2009) showed that Andrew had a moderately low score (standard score= 78, percentile rank =7) and Craig had an average score (standardised score= 90, percentile rank =26). Non-verbal IQ was also tested using the Raven's Coloured Progressive Matrices (Raven, Raven and Court, 1998), showing that both children were in the 75th percentile (Grade II "definitely above average intellectual capacity").

Probes/Speech Measures

An untreated wordlist targeting each child's specific lingual errors was recorded at each assessment time point. Target wordlists were selected based on information from the referring SLT (Andrew: /n/, Craig: velars stops). Untreated wordlists consisted of 36 words (Table 2). Andrew's wordlist contained 39 tokens of /n/ and Craig's wordlist contained 41 tokens of velars, including velar plosives and nasal stops. These words were never used in the course of therapy, allowing us to check for generalisation of targets. Each wordlist contained the "in error" lingual targets (denoted by underlining in table 2) in (all singleton) word initial, (mostly intervocalic) medial and (mostly singleton) final positions in a variety of vowel environments.

Insert table 2 about here

A narrow phonetic transcription of the probe (whole words) from the first block of therapy was performed by the treating clinician (first author) using the acoustic and lip-camera data. Post therapy block one, acoustic, ultrasound and lip-camera data were used for narrow phonetic transcription of the probe data, also carried out by the treating clinician. Broad phonetic transcriptions were carried out by the remaining authors using audio data only, in order to compare

PTCC scores. The order of the sessions was randomised so that authors two and three were blinded to information about which time point the data was derived from. However, this was not possible for the treating clinician. From these transcriptions we calculated percent target consonant correct (PTCC) at each time point by giving each token a score of 1 (correct) or 0 (incorrect). PTCC listener agreement was calculated. PTCC scores are based on data from single words only, not from a connected speech sample. PTCC scores from the treating clinician are presented, with inter-rater reliability reported from the remaining two authors' PTCC scores.

Therapy

Therapy was provided by a qualified speech and language therapist/pathologist (SLT, the first author). The first block of therapy used Articulatory Animations (AAs) from the iPad (Apple 2012) app Speech Trainer 3D (Smarty Ears, 2011) as a VAM and the second block of therapy used Ultrasound Visual Biofeedback (U-VBF). Since this was a pilot study with only two speakers it was not possible to randomise to groups. We conducted the VAM therapy first since we initially expected this to have less of an effect on speech outcomes than U-VBF for which the current evidence base is stronger. See Roxburgh, Scobbie & Cleland (2015) for details regarding therapy.

Multiple Listener Perceptual Evaluation

The perceptual evaluation is a modification of a two-alternative forced choice experimental design, using data from the untreated wordlist probes (Table 2). The audio materials were extracted from the untreated wordlists. Comparisons were between pairs of tokens of the same word drawn from two different time points during therapy (assessment sessions one to six). Listeners were told they would hear two versions (first V1 then V2) of the same single real target word. The order of presentation was counterbalanced, so that either V1 or V2 could be chronologically earlier (A) or chronologically later (B) in therapy. Listeners were asked to decide which acoustic stimulus sounded "closer to the English target word" presented orthographically on the screen.

This design requires a listener to holistically evaluate the two stimuli relative to each other *and* against their expectations of production norms. We expect the listeners to use a mix of segmental phonological and phonetic (including voice and prosodic) characteristics of the speech as the basis of their judgement, and we gave no instructions on what balance to use. Listeners were, however, provided with explicit information regarding therapy targets, to encourage them to focus on the target phoneme, a focus reinforced implicitly by the frequent occurrence of the target in the wordlist. We expect, but cannot guarantee, that a segmental phonological comparison to the abstract target would be given greater weight when one of the audio stimuli sounded phonologically correct for the target segment, and one incorrect. On the other hand, if both audio stimuli are phonologically similar (both correct, or both wrong), we would expect listeners to be more likely to base their judgement on fine phonetic differences or prosodic differences between the stimuli. If both stimuli are indistinguishable we expect chance-level scores. In none of these cases the comparisons require phonetic training or specialised knowledge. And, by using a bank of listeners, a significant result will only occur if enough of them concur in their judgement, at a level greater than chance.

Data was presented using PRAAT (Boersma and Weenink, 2013). For each stimulus pair of tokens, listeners were allowed to listen up to three times at their own pace. In our main statistical analysis (see below), we considered each listener as independent, and test each speaker's overall session-to-session judgement, based ultimately on the cumulative weight of the binary judgements of the stimuli pairs. If the listener showed an overall statistically-significant preference for the chronologically later session B, this was interpreted as an indication of an improvement in the client's speech associated with the therapy target which is present in all the pairs. This, we assume, will reflect more general improvements in accuracy, intelligibility and/or acceptability, for that listener. If enough listeners showed a significant difference in the same direction, for this trend in

agreement to be significant, we report a change in accuracy. We also quantify the consistency of their judgement using Fleiss's kappa.

This paper reports on a situation in which listeners were fully familiar with the protocol and fully practiced in the method. This is because all judgments reported here come from a listener's second full experience of this task, approximately one month after listening to a different child. The time gap was designed to avoid priming. In fact, prior experience consisted of the same listening experiment, undertaken on the other speaker (either Andrew or Craig, as appropriate). A full analysis of the listeners' first session and of the relationship between these two independent evaluations of Andrew and Craig will be presented elsewhere.

In order to avoid over-familiarity with the individual probe tokens, we avoided comparing every possible time point in every possible permutation from each of the six assessment sessions outlined in Table 1. In fact, even the BL-M comparison uses different sessions to the other two comparisons, so they could all be viewed as independent. We compare:

a. VAM (pre/post) Comparison: immediately before and after therapy block one using VAM (session 2 vs. session 3)
b. U-VBF (pre/post) Comparison: immediately before and after therapy block two using U-VBF (session 4 vs. session 5)
c. BL-M Comparison: baseline to maintenance (session 1 vs. session 6)

Individual words were edited from longer recordings (three words per recording) hence silence was included either side of each word where possible. An additional 0.5 second silence was presented between V1 and V2 as the inter-stimulus interval. The number of tokens for each comparison block was 36 single words, giving a total of 108 comparisons. As listeners could listen to each comparison up to three times, this meant that they listened to at least 108 and at most 324 pairs for comparison: we did not observe what listeners did. The order of the words was uniquely randomised for each listener. The time taken for listeners to complete the task was approximately 30 minutes, so the rate

of exposure was only about 3.5 pairs per minute. Explicit but optional rest breaks were provided every 18 tokens in the PRAAT script to reduce listener fatigue.

Analysis

A non-parametric sign test (Corder and Foreman, 2014) was used for statistical analysis of each of a listener's three blocks of comparisons. The number of words from session B (chronologically later in time) judged to be more target-like for each individual listener was tested for significance at p<.05. For a two-tailed test of 36 lexical pairs, this requires a listener to assign more than 25 to one or other category (*p*=.029). Since it can be argued that the three comparison blocks are not independent, we will also report significance at p/3 as a Bonferroni adjustment, i.e. significance is set at *p*<.017 (a threshold of 26/36, *p*=.011). Listener agreement was then calculated, based on the overall number of session A or session B selected (including non-significant preferences) and was statistically tested using a Fleiss' Kappa (Fleiss, 1981) for each word-pair comparison. A Fleiss' Kappa was also calculated for WI, WM and WF contexts. A Fleiss' Kappa can be used to measure agreement among listeners and transcribers (see below). Fleiss' Kappa results can be interpreted in the following way: < .40 = Poor agreement; .60 – .74 = Intermediate to good agreement; ≥ .75 = Excellent agreement (Fleiss, 1981). The level of consistency among the listeners, combined with the number of listeners exhibiting a Bonferroni-corrected significant preference for improvements gives in our view a robust, conservative and replicable rating.

Listener responses can also be compared to Percent Target Consonant Correct (PTCC) for the relevant sessions, which were derived from phonetic transcriptions of the audio samples. Consistency of PTCC for three transcribers is also reported using Fleiss' kappa. While a direct correlational analysis is not possible due to the differing methodologies, some qualitative remarks on the two approaches to testing for improvement can be made.

Results

Craig: Phonetic Transcriptions

Figure 1 shows Craig's PTCC scores from all three transcribers over time. Transcriber 1 (first author and treating clinician) gave Craig a PTCC score of 22%, which remained relatively stable, although slightly higher, in the pre-VAM assessment with a score of 26%. In the Post-VAM assessment Craig's PTCC had increased to 76%, with correct productions of [ŋ], and velar plosives in all word positions. Scores remained stable over the inter-therapy break. After the second block of therapy this had risen to 93% in the Post-U-VBF session, which remained relatively stable, although slightly lower, at 90% in the maintenance session. All three transcribers agreed on individual token pairs over 70% of the time across all six assessment sessions (mean = 76% range = 71%-85%, "intermediate to good agreement"). Statistical analysis showed that the highest agreement across transcribers was found in the Pre-UTI session (Fleiss' Kappa = .7375) with all three transcribers agreeing on 33/41 tokens. The lowest agreement was found in the maintenance session (Fleiss' Kappa = .5597) with all three transcribers agreeing on 29/41 tokens.

Insert figure 1 about here

Craig: Perceptual Evaluation

The number of B (chronologically later in time) selected within each comparison was calculated. Overall, in the VAM Comparison, listeners selected B for 326/432 token pairs (75%) suggesting that Craig's productions of velars in single words post-therapy were closer to the target 75% of the time. In the U-VBF Comparison, B was selected for 249/ 432 tokens (58%), i.e. post-therapy recording were selected 58% of the time, and in the BL-M Comparison, 350/432 (81%) showing that productions from the maintenance session were selected 81% of the time.

Statistical analysis of the listener responses using a two-tailed sign test, with a chance level of $p < .05$ was calculated. Results showed that almost all listeners selected B significantly more than A within

the VAM Comparison (11/12, or 10/12 when Bonferroni adjusted). In the U-VBF Comparison, only one listener (Listener 20) selected B significantly more than A (and in fact at p= .0288, L20 was not significant when Bonferroni adjusted). Two listeners (14 and 13) selected A more than B (denoted in Table 3 by boldface), however neither were significant. Overall there was significant improvement in the overall BL-M Comparison, with all listeners selecting B significantly more than A, even after Bonferroni adjustment. Table 3 shows individual listener results.

Insert table 3 about here

Listener agreement was calculated for each word pair within each comparison (VAM, U-VBF and BL-M), also with Kappa. Results show that 100% listener agreement was found in 27/36 words in at least one comparison. Fleiss' Kappa results show that there was agreement between listeners for all word positions in all three comparisons, with the lowest agreement found in WI position in the VAM Comparison (Fleiss' Kappa = .0411) and the highest agreement found in WF position in the BL-M Comparison (Fleiss' Kappa = .5927). Based on Fleiss' (1981) interpretation of results however, both of these results demonstrate poor agreement between listeners on a word-by-word basis.

Craig: Results Summary

Listeners in the perceptual evaluation selected B (chronologically later in time) more than A (chronologically earlier in time) within all three comparisons (VAM, U-VBF and BL-M). This corresponds with the mean PTCC scores from transcribers, with higher PTCC scores found in the same sessions selected as being "closer to the English target". All three transcribers reported an increase in PTCC in the VAM comparison (pre-VAM 24% PTCC to post-VAM 84% PTCC), the U-VBF comparison (pre-U-VBF 80% to post-U-VBF 93% PTCC) and the BL-M comparison (22% PTCC at baseline to 85% PTCC in the maintenance session).

Andrew: Phonetic Transcriptions

Figure 2 shows Andrew's PTCC scores from all three transcribers over time. At baseline he achieved a PTCC score of 5% (transcriber 1, first author and treating clinician), which remained relatively stable, although slightly higher, in the pre-VAM assessment (8%). In the Post-VAM assessment Craig's PTCC had increased to 21%. Scores showed an increase in the inter-therapy break, increasing by 10 percentage points to 31% in the Pre-U-VBF assessment session. After the second block of therapy this had decreased to 5% in the Post-U-VBF session, with an increase to 21% in the maintenance session. Results showed that all three transcribers agreed the majority of the time (mean=72% range=59%-80%). Statistical analysis showed that the highest agreement across transcribers was found in the Maintenance session (Fleiss' Kappa = .6538 "intermediate to good") with all three transcribers agreeing on 31/39 tokens. The lowest agreement was found in the pre-VAM session (Fleiss' Kappa = .0969), despite all three transcribers agreeing on 30/39 tokens.

Insert figure 2 about here

Andrew: Perceptual Evaluation

Overall, in the BL-M Comparison, listeners selected B for 254/360 comparisons (71%). In other words 71% of Andrew's productions in the maintenance session were judged as closer to the target, suggesting improvement. In the VAM Comparison, B was selected for 228/360 tokens (63%), showing the post-therapy productions were selected more than pre-therapy, and in the U-VBF Comparison, 151 /228 (42%), unexpectedly showing that pre-therapy productions were selected more than post-therapy productions.

Statistical analysis showed that within the VAM Comparison, all listeners selected B (chronologically later in time) as "closer to the English target", but only two listeners selected B statistically significantly more than A (Listener 17, p=.0039; Listener 11, p=. 0113, both still significant at the adjusted threshold of p<.017). Within the U-VBF Comparison, a very different trend was observed, in which 7/10 listeners selected A (chronologically *earlier* in time) as "closer to the English target" more often than B (denoted by boldface in table 4), though only one (Listener 18) selected A significantly

more than B (p <.0001). In the BL-M Comparison, all listeners selected B more than A, with 7/10 listeners producing significant results, or 6/10 with Bonferroni adjustment. Results for each individual listener are outlined in Table 4. Overall, listeners selected B more often than A, but more tokens of A were selected in the U-VBF Comparison, showing that although there was an overall improvement after both types of therapy, there was statistically no change, and perhaps even a trend indicating a slight deterioration, after the second block of therapy with ultrasound.

Insert table 4 about here

Results show that 100% listener agreement was found in 17/36 words in at least one comparison. Fleiss' Kappa results show that there was general agreement between listeners for all word positions in all three comparisons, with the lowest agreement found in WI position in the VAM Comparison (Fleiss' Kappa =.0770) and the highest agreement found in WF position in the U-VBF Comparison (Fleiss' Kappa =.4351). Again both of these results demonstrate poor agreement between listeners, based on Fleiss (1981) interpretation.

Andrew: Results Summary

Results show that when listeners selected B (chronologically later) more than A (chronologically earlier in time) in the VAM and BL-M Comparisons, phonetic transcriptions also showed an increase in PTCC in these comparisons. In the U-VBF Comparison, when listeners selected A more than B, PTCC also decreased in the U-VBF post-therapy session, so both measures indicate a slight deterioration in this block. The transcribers showed only 59% agreement in the U-VBF comparison, suggesting that the data was more difficult to interpret than in the other sessions, Fleiss' Kappa results show that the lowest listener agreement was in WI position in VAM Comparison, with the Fleiss' Kappa also showing lowest transcriber agreement in the pre-VAM session.

Discussion

This study piloted a new perceptual methodology for evaluating changes after an intervention. Previous literature for CP speech suggests, for example, using two listeners with a large amount of experience in the field for phonetic transcription (Sell, 2005 and other references above). We sought to determine via a multi-listener perceptual evaluation whether listeners are able to detect any improvement in production of untreated single words presented as audio stimuli, in comparison to the target English word and with knowledge of the goal of therapeutic intervention. We then compared this to PTCC scores, derived from phonetic transcriptions by three experienced phoneticians.

Our initial aim was to determine whether listeners select time point B (chronologically later in the therapy period) as closer to the English target more often than time point A (chronologically earlier in the therapy period), hence indicating post-therapy improvement. Since our therapy design compared two interventions, and since probes had been made in six sessions, we were able to evaluate our perceptual method in the comparison of three different pairs of time points- before and after therapy with VAMS, before and after U-VBF and overall improvement from baseline to maintenance. Our hypothesis was confirmed with the majority of later time-points being selected as "closer to the English target" by most listeners and especially for "Craig". In fact, listeners detected a further improvement in Craig's speech after ultrasound despite an early improvement after the first block of VAM therapy. Clinically, this is to be interpreted with caution because he had clearly acquired the new speech sound before commencing the U-VBF therapy. Indeed, VBF is thought to be most useful for establishing motor programmes for new articulations (Gibbon and Wood, 2010), thus probably rendering it unnecessary once Craig had learned to produce a velar articulation in the VAM block of therapy.

For Andrew, listener judgements unexpectedly indicated a decrease in intelligibility/accuracy post-therapy using ultrasound (U-VBF Comparison), with seven out of 10 listeners selecting A more

than B. This is contrary to previous studies reporting success with U-VBF (Bacsfalvi, Bernhardt and Gick, 2007; Bacsfalvi, 2010; Bacsfalvi and Bernhardt, 2011; Cleland et al., 2015), highlighting the need to design larger studies which compare U-VBF with competing therapies, rather than no treatment.

Secondly, we sought to determine whether our perceptual experiment was in line with PTCC from experienced phoneticians. For both children, phonetic transcription showed an increase in percentage of targeted consonants correct from initial baseline to maintenance, three months after therapy ceased. For Andrew this improvement was modest, rising from 5% PTCC at baseline to only 21% PTCC at maintenance. This is unlikely to represent a clinically significant improvement in Andrew's production of /n/ suggesting that neither therapy was particularly effective. In contrast, Craig improved from 22% PTCC at baseline to 90% PTCC at his maintenance recording, suggesting he had successfully integrated velars into untreated words.

Despite previous literature suggesting that transcriptions from single transcribers are unreliable and multi-listener judgements being preferable (Kuehn and Moller, 2000; Lohmander and Olsson, 2004; Britton et al., 2014) results of our new methodology closely corroborate the phonetic transcription. We suggest that when listeners select B (chronologically later, i.e. post-therapy) as "closer to the English target", the PTCC score also increase. However, with the differing methodologies of the perceptual evaluation and the phonetic transcriptions it was not possible to correlate results statistically. Previous literature suggests that point-by-point reliability for broad phonetic transcription is often in the 90-95% range and for narrow transcription is often around 80% (Shriberg & Lof 1991; Shriberg, Austin, Lewis, McSweeny & Wilson, 1997). Preston, Ramsdell, Oller, Edwards & Tobin (2011) point out that it is more difficult to achieve agreement on disordered speech, with complex speech disorders such as those found in cleft palate often being associated with low inter-rater agreement (Shriberg & Lof, 1991). Gooch, Hardin-Jones, Chapman, Trost-Cardamone & Sussman (2001) found an average of 40% agreement across listeners (range 19%-71%) when comparing listener judgements against transcriptions of compensatory articulations. Based on

the range of reliability proposed by Shriberg & Lof (1991) and Shriberg et al (1997), results would suggest that our average of 74% accuracy across both speakers is not reliable, highlighting the need for multiple listener perceptual evaluations such as this.

Our perceptual evaluation provides a quick and easy method of testing pre- and post-therapy speech with multiple-listeners. From a practical perspective, an MFC document in PRAAT version 5.3.57 (Boersma, and Weenink, 2013), can be modified by the research team by copying and pasting audio file names into the document. This process is quick and easy and takes no longer than 60 minutes. Since conducting the study, three small projects have replicated the methodology with ease (Alexander, 2015; Thompson, 2015 and Young, 2015).

Although the current study used phonetically trained listeners who have experience in listening to disordered speech, the methodology is designed so that lay listeners can also be used. Listeners were asked to select which version of a word was “closer to the English target” based on their own phonological and phonetic intuitions, which does not require phonetic skill. Future studies should compare ratings by expert and lay listeners. Furthermore, although no obvious differences were found in the current study between the PTCC scores and the perceptual evaluation, in a master’s thesis using the same methodology to evaluate pre, during and post-therapy changes in a child with Childhood Apraxia of Speech, listeners identified subtle improvements between mid-therapy recordings and post-therapy recordings, both of which were rated as 100% on target by a transcriber (Young, 2015). Results from Young (2015) suggest that this perceptual evaluation method might be useful for detecting subtle, improvements in acceptability, without the need for time-consuming narrow transcription.

In terms of agreement between listeners, statistical analysis showed varied level of agreement, with listener judgements matching PTCC scores derived from phonetic transcriptions. Listener judgements were more reliable for Craig than Andrew, which is probably the result of chance levels (i.e. guessing) when tokens from different time-points were indistinguishable, and in this sense the perceptual evaluation is quite different from a phonetic transcription which is not designed to detect

improvement without further analysis (for example calculating percentage consonants correct). Some productions may also have been ambiguous, with double articulations suspected through narrow phonetic transcriptions and confirmed by ultrasound analysis (not reported here). Whilst it may appear from statistical analysis that there was poor agreement between listeners and at times between transcribers, it should be noted that the kappa is a conservative statistical measure as it assumes a high level of agreement obtained by chance when judgements were not evenly distributed (Cordes, 1994; Brunnegard & Lohmander, 2007). Perhaps the low kappa in a situation with high agreement is due to a ceiling effect.

Limitations and Future Research Implications

In this study the perceptual evaluation was piloted only on phonetically trained listeners. This had the advantage that it was straightforward to explain to listeners which phoneme they should focus on in the stimuli. However, previous literature states the benefits of using lay listeners in perceptual evaluations of CP speech. It would therefore be beneficial to further test this methodology on lay listeners to compare with phonetically trained listener responses, adding further to inter-rater reliability. Using lay-listeners would also have the advantage that is might be possible to employ this methodology using remote listeners via the internet or even using Crowdsourcing. Byun, Halpin, and Szeredi, (2015) found that it was possible to use lay-listeners to rate the speech of children with mild articulatory difficulties (/ɹ/ misarticulations) using the Crowdsourcing platform "Amazon Mechanical Turk" quickly and easily. Using our method, it would be possible to do the same for more severe SSDs, especially where speech is less intelligible.

Conclusion

In conclusion, the perceptual evaluation shows promise as a method of evaluating speech outcomes from any speech therapy such as ultrasound and visual articulatory models. In this case, the evaluation showed substantially improved speech in one speaker and little gain for the other,

with results mirroring the PTCC scores despite weak listener agreement. Although further testing is needed, it should be possible to extend the method to other subgroups of speakers with SSD.

Acknowledgements

Thanks are due to the participants: both the children and the phonetically trained listeners, for participating in the study. Thanks are also due to Gillian Cairns for assisting in recruitment of children, Eleanor Lawson for providing assistance with PRAAT, Robert Rush and Felix Schaeffler for statistical assistance and Steve Cowen and Alan Wrench for technical assistance.

Declaration of Interest

This study was supported by a PhD Bursary at Queen Margaret University from 2011-2015.

References

Alexander, K. (2015). Effectiveness of ultrasound visual biofeedback therapy. *Unpublished Honours Project,* Queen Margaret University, Edinburgh.

Amorosa, H., von Benda, U., Wagner, E., & Keck, A. (1985). Transcribing phonetic detail in the speech of unintelligible children: A comparison of procedures. *British Journal of Disorders of Communication,* 16 (3), 281-287.

Apple. (2012). *The new iPad.* [online] Available at: http://store.apple.com/uk/browse/home/shop_ipad/family/ipad [Accessed March 16 2012].

Articulate Instruments. (2013). *Articulate Assistant Advanced softwareTM Ultrasound Module User Manual*. Revision 215, Articulate Instruments Ltd.

Bacsfalvi, P. (2010). Attaining the lingual components of /r/ with ultrasound for three adolescents with cochlear implants. *Canadian Journal of Speech-Language Pathology & Audiology,* 34, 206-217.

Bacsfalvi, P., Bernhardt, B.M., & Gick, B. (2007). Electropalatography and ultrasound in vowel remediation for adolescents with hearing impairment. 4th International Electropalatography (EPG) Symposium, held in Edinburgh in September 2005. *Advances in Speech Language Pathology,* 9, 36-45.

Bacsfalvi, P., & Bernhardt, B.M. (2011). Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography. *Clinical Linguistics & Phonetics,* 25 (11/12), 1034-1043.

Bessell, A., Sell, D., Whiting, P., Roulstone, S., Albery, L., Persson, M., & Ness, A. R. (2013). Speech and language therapy interventions for children with cleft palate: a systematic review. *The Cleft Palate-Craniofacial Journal,* 50(1), e1-e17.

Boersma, P., & Weenink, D. (2013). *PRAAT doing phonetics by computer.* Version 5.3.57. [online]. Available at: www.praat.org. Accessed on: March 2014.

Bressmann, T., Radovanovic, B., Kulkarni, G. V., Klaiman, P., & Fisher, D. (2011). An ultrasonographic investigation of cleft-type compensatory articulations of voiceless velar stops. *Clinical Linguistics & Phonetics,* 25(11-12), 1028-1033.

Britton, L., Albery, L., Bowden, M., Harding-Bell, A., Phippen, G., & Sell, D. (2014). A Cross-Sectional Cohort Study of Speech in Five-Year-Olds with Cleft Palate +/- Lip to Support Development of National Audit Standards. Benchmarking Speech Standards in The United Kingdom. *The Cleft Palate-Craniofacial Journal,* 51(4), 431-451.

Brunnegard, K., & Lohmander, A. (2007). A Cross-Sectional Study of Speech in 10-Year-Old Children With Cleft Palate: Results and Issues of Rater Reliability. *Cleft Palate–Craniofacial Journal*, 44 (1), 33-44.

Buckingham, H.W., & Yule, G. (1987). Phonemic false evaluation: Theoretical and clinical aspects. *Clinical Linguistics*, 1, 113–125.

Byun, T. M., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of communication disorders*, 53, 70-83.

Cleland, J., Scobbie, J.M., Nakai, S. & Wrench (2015). Helping Children Learn Non-native Articulations: The Implications for Ultrasound-based clinical Intervention. *Proceeding of the International Congress of Phonetic Science (0698),* 1-5.

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical linguistics & phonetics,* 0, 1-23.

Corder, G.W., & Foreman, D.I. (2014). *Non-Parametric Statistics: A Step-by-Step Approach. (2nd Edition).* Hoboken, New Jersey: Wiley.

Cordes AK. (1994). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech & Hearing Research*, 37, 264–78.

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2002). *Diagnostic Evaluation of Articulation and Phonology.* London: The Psychological Corporation.

Dunn, L.M., Dunn, D.M., & National Foundation for Educational Research. (2009). *British Picture Vocabulary Scale 3rd ed. (BPVSIII).* Great Britain: Wascana Ltd partnership and GL Assessment.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Gibbon, F.E. & Wolters, M. (2005). A new application of ultrasound to image tongue behaviour in cleft palate speech. *Poster presentation at the Craniofacial Society of Great Britain and Ireland Annual Scientific Conference,* Swansea, UK (13-15 April 2005).

Gibbon, F.E. & Wood, S. (2010). Visual feedback therapy with electropalatography (EPG) for speech sound disorders in children. In L. Williams, S. McLeod and R. McCauley (Eds.) *Interventions in Speech Sound Disorders.* Brookes: Baltimore. Pp. 509-536.

Gooch, J. L., Hardin-Jones, M., Chapman, K. L., Trost-Cardamone, J. E. and Sussman, J. (2001). Reliability of listener transcriptions of compensatory articulations. *Cleft Palate Journal*, 38, 59–67.

Harding, A., & Grunwell, P. (1998). Active versus passive cleft-type speech characteristics. *International Journal of Language and Communication Disorders,* 33(3), 329-352.

Kuehn, D., & Moller, K.T. (2000). Speech and Language issues in the cleft palate population: the state of the art. *The Cleft Palate-Craniofacial Journal,* 37, 348-348.

Lohmander A., & Olsson, M. (2004). Methodology for Perceptual Assessment of Speech in Patients with Cleft Palate: A Critical Review of the Literature. *The Cleft Palate-Craniofacial Journal,* 41,.64-70.

McAllister Byun, T., Hitchcock, E.R., & Swartz, M.T. (2014). Retroflex Versus Bunched in Treatment for Rhotic Misarticulation: Evidence From Ultrasound Biofeedback Intervention. *Journal of Speech, Language and Hearing Research,* October, 1-15.

Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology,* 22(4), 627-643.

Preston, J.L., Ramsdell, H.L., Oller, D.K., Edwards, M.L., & Tobin, S.T. (2011). Developing a Weighted Measure of Speech Sound Accuracy. *Journal of Speech, Language, and Hearing Research*, 54, 1–18.

Rahimov, F., Jugessur, A., & Murray, J.C. (2012). Genetics of Nonsyndromic Orofacial Clefts*. The Cleft Palate-Craniofacial Journal,* 49(1), 73-91.

Raven, J., Court, J. & Raven, J. (1998). *Raven's Progressive Matrices and Raven's Coloured Matrices*. Oxford, England: Oxford Psychologists Press.

Roxburgh, Z., Scobbie, J.M. & Cleland, J. (2015) Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback. *Proceedings of the 18th ICPhS, Glasgow (0858),* 1-5.

Royal College of Speech and Language Therapists (RCSLT). (2005). *RCSLT Clinical Guidelines.* Oxon: Speechmark Publishing.

Ruscello, D., & Vallino, L. (2014). The Application of Motor Learning Concepts to the Treatment of Children with Compensatory Speech Sound Errors. *SIG 5 Perspectives on Speech Science and Orofacial Disorders,* 24(2), 39-47.

Sell, D. (2005). Issues in perceptual speech analysis in cleft palate and related disorders: a review. *International Journal of Language and Communication Disorders,* 40(2), 103-121.

Semel, E., Wiig, E.H., & Secord, W.A. (2006). *Clinical Evaluation of Language Fundamentals. 4th ed*. London: Pearson Assessment.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research,* 40, 708–722

Shriberg, L.D., & Lof, G.L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, 1, 171-189.

Smarty Ears. (2011). *Speech Trainer 3D.* [online] Available at: http://smartyearsapps.com [Accessed November 2011].

Thompson, E. (2015). Intelligibility pre/post therapy in children with cleft lip and palate. *Unpublished Honours Project,* Queen Margaret University, Edinburgh.

Trost, J. E. (1981). Articulatory additions to the classical description of the speech of persons with cleft palate. *Cleft Palate Journal,* 18, 193–203.

Vallino-Napoli, L. (2011). Evaluation and Evidence-Based Practice. In: Howard, S., and Lohmander, A. (Eds) *Cleft Palate Speech: Assessment and Intervention (pp. 317-358).* Chichester: Wiley-Blackwell.

Young, E. (2015). The Effectiveness of Ultrasound Visual Biofeedback in the remediation of an Idiosyncratic Case of Alveolar Backing. *Unpublished Master's Thesis,* Queen Margaret University, Edinburgh.

Zharkova, N. (2013). Using ultrasound to quantify tongue shape and movement characteristics. *The Cleft Palate-Craniofacial Journal,* 50, 76-81.

Ziegler, W., & Zierdt, A. (2008). Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders*, 41,553–577.

Week 1 Assessment 1	Week 2 Assessment 2	Weeks 3-10	Week 11 Assessment 3	Week 16 Assessment 4	Weeks 17-23	Week 24 Assessment 5	+3 Months Assessment 6
Baseline	Pre VAM	Therapy Block 1: Speech Trainer 3D (VAM)	Post VAM	Pre U-VBF	Therapy Block 2: U-VBF	Post U-VBF	Maintenance
DEAP (phonology) Untreated Wordlist	DEAP (phonology) Untreated Wordlist		DEAP (phonology) Untreated Wordlist	DEAP (phonology) Untreated Wordlist		DEAP (phonology) Untreated Wordlist	DEAP (phonology) Untreated Wordlist

Table 1 Probe and treatment schedule (*Diagnostic Evaluation of Articulation and Phonology (DEAP, Dodd, Hua, Crosbie, & Holm, 2002)) N.B Different shades of grey denote the three blocks for comparisons

Craig: Velars	Andrew: /n/
WI	**WI**
car, comb, computer	nachos, gnome, nappy
cup, guitar, goat	knot, necklace, nuts
carrots, gum, gate	neeps*, kneeling, knitting
cage, gorilla, gas	nibbling, notebook, nose
WM	**WM**
lego, magnet, sugar	lemonade, vanilla, sunny
nuggets, cookie, jacket	dinner, dinosaur, funny
necklace, bucket, singer	onions, brownie, tuna
banging, angry, kangaroo	banana, piano, animals
WF	**WF**
smoke, snack, flag	garden, leprechaun, can
magic , snowflake, jog	phone, violin, green
warthog, handbag, strong	snowman, medicine, popcorn
ring, skiing, jumping	bone, skeleton, curtain

Table 2 Speech Stimuli: Untreated Wordlists (* Scottish word for turnip)

Listener Number	VAM Comparison	U-VBF Comparison	BL-M Comparison
1	.0003**	1.1321	<.0001***
2	.0003**	.4050	<.0001***
3	.0003**	.8679	<.0001***
4	.1325	.8679	.0113**
5	.0039**	.1325	.0003**
6	.0288*	.1325	.0003**
13	<.0001***	**.8679**	.0003**
14	.0039**	**.6177**	<0.0001***
20	.0113**	.0288*	.0113**
21	.0113**	.6177	.0003**
22	.0003**	.0652	.0012**
25	.0113**	.0652	<0001***

Table 3 Sign Results for Craig (NB *p < .05. **p < .01. ***p <.001)

Listener Number	VAM Comparison	U-VBF Comparison	BL-M Comparison
7	.4050	**.2430**	.0012**
9	.2430	**.4050**	.1325
10	.6177	.8679	.0012**
11	.0113**	**.1325**	.0652
12	.0652	**.4050**	.0288**
15	.0652	**.4050**	.0113**
16	.1325	.6177	.0113**
17	.0039**	**.6177**	.0113**
18	.4050	**<.0001*** **	.2430
19	.6177	1.1321	.0113**

Table 4 Sign Results for Andrew for all three comparisons (p* < .05. ***p* < .01. ****p* <.001)

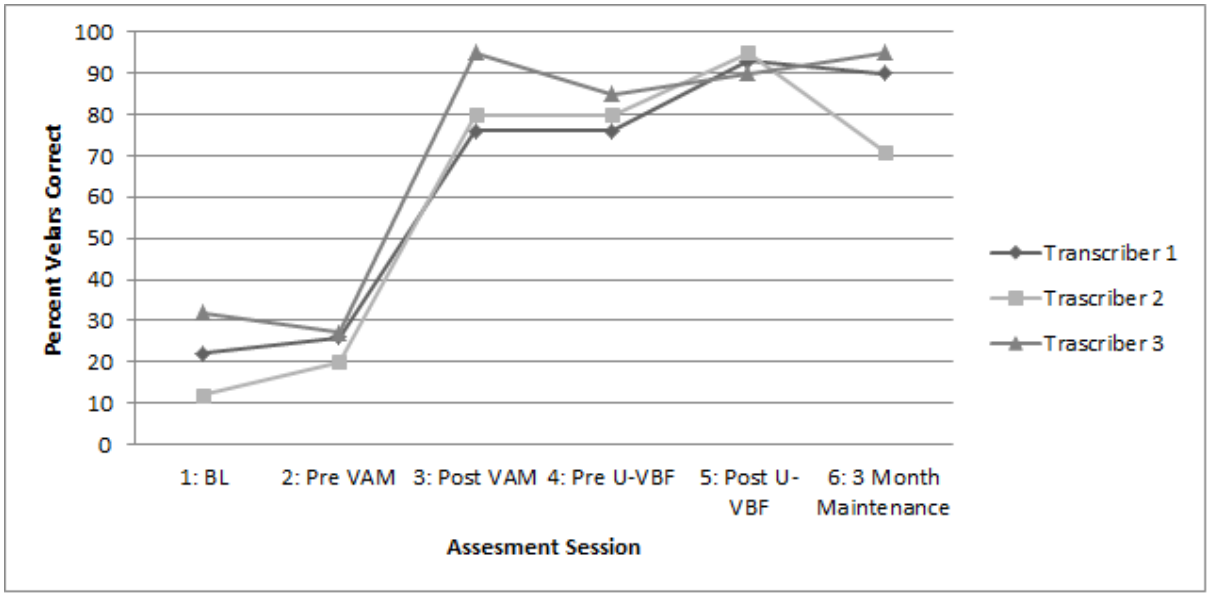

Figure 1 Craig's PTCC Scores from all three transcribers for Untreated Wordlist

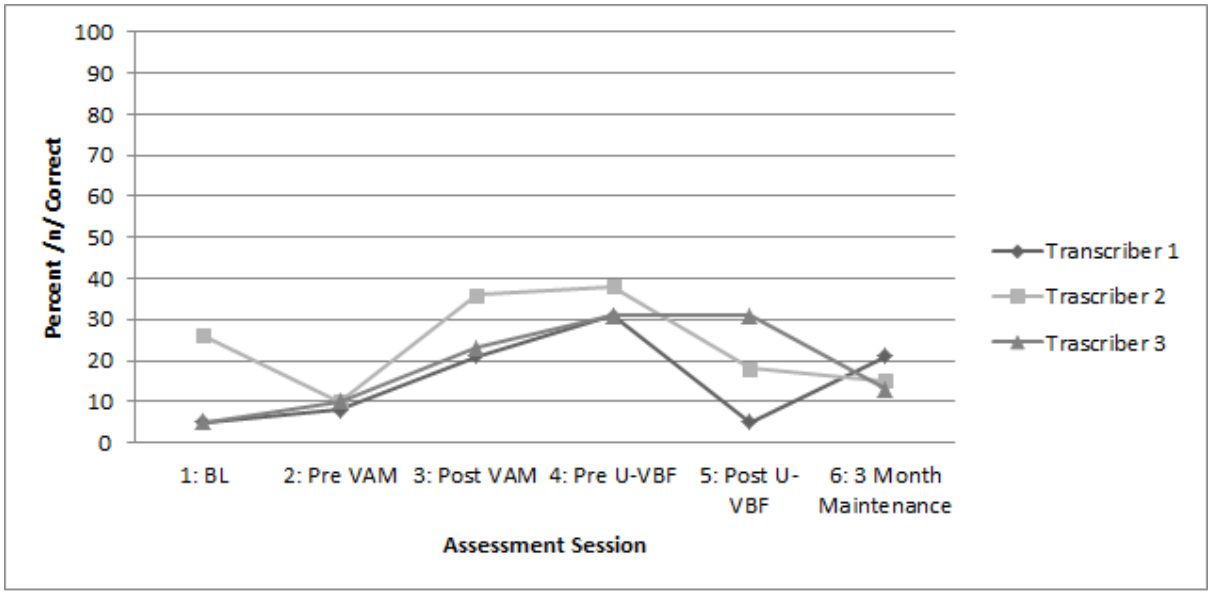

Figure 2 Andrew's PTCC Scores from all three transcribers for Untreated Wordlist